Title

stata.com

Example 8b — Random effects, endogenous covariate, and endogenous sample selection

Description Remarks and examples Also see

Description

In [ERM] **Example 8a**, we ignored the observations that were dropped because of missing data on GPA. In this example, we show you how to fit a model with a continuous outcome, a continuous endogenous covariate, endogenous sample selection, and random effects.

Remarks and examples

stata.com

In the last example, the researchers excluded students who dropped out of college because they are missing college GPA on these students. Thus they were estimating the parameters for only the population of students who graduate from college. Now let's suppose that the researchers are interested in the expected college GPA for all the students who enrolled, even those who dropped out. What would their GPA be if they had remained in school?

The researchers assumed that unobserved student ability affected both college and high school GPAs. They also suspect that unobserved ability affects the decision to stay in school, so they could have an endogenously selected sample. The researchers have data on whether the students have participated in a retention program (`program`) and whether they had a roommate from the same college (`roommate`). They use these variables in addition to high school GPA and parent's income to model whether the student graduates.

The researchers assumed that there are unobserved characteristics of the college that affects college GPA. They also assumed that unobserved college characteristics such as the availability and type of extracurricular activities and the rigor of the curriculum affect whether the students graduate. They account for these unobserved college-level factors that may affect the probability of graduating and the final college GPA of the students by including random effects in both of these equations.

```
. xteregress gpa income, endogenous(hsgpa = income i.hscomp, nore)
> select(graduate=hsgpa income i.roommate i.program)
(setting technique to bhhh)
Iteration 0:  Log likelihood = -750.88823
Iteration 1:  Log likelihood = -750.14312
Iteration 2:  Log likelihood = -750.09077
Iteration 3:  Log likelihood = -750.03772
Iteration 4:  Log likelihood = -750.03525
Iteration 5:  Log likelihood = -750.03163
Iteration 6:  Log likelihood = -750.03143
Iteration 7:  Log likelihood = -750.03082
Iteration 8:  Log likelihood = -750.03081
Iteration 9:  Log likelihood = -750.03069
(switching technique to nr)
Iteration 10: Log likelihood = -750.03068
Iteration 11: Log likelihood = -750.03067
```

```
Extended linear regression                      Number of obs     =      2,000
                                                        Selected  =      1,372
                                                    Nonselected   =        628

Group variable: collegeid                       Number of groups  =        100

                                                Obs per group:
                                                              min =         20
                                                              avg =       20.0
                                                              max =         20

Integration method: mvaghermite                 Integration pts.  =          7

                                                Wald chi2(2)      =    2498.14
Log likelihood = -750.03067                     Prob > chi2       =     0.0000
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
gpa						
income	.0580659	.0039428	14.73	0.000	.0503381	.0657938
hsgpa	.975956	.0719006	13.57	0.000	.8350334	1.116879
_cons	-.7193664	.2132662	-3.37	0.001	-1.13736	-.3013723
graduate						
hsgpa	1.59638	.5058428	3.16	0.002	.604946	2.587813
income	.2111094	.0256101	8.24	0.000	.1609145	.2613043
roommate						
Yes	1.16331	.0893901	13.01	0.000	.9881092	1.338512
1.program	.8719825	.0858947	10.15	0.000	.7036319	1.040333
_cons	-6.488787	1.512468	-4.29	0.000	-9.453169	-3.524405
hsgpa						
income	.0487467	.0016938	28.78	0.000	.0454269	.0520664
hscomp						
Moderate	-.1594138	.0121475	-13.12	0.000	-.1832225	-.1356051
High	-.2532709	.0195334	-12.97	0.000	-.2915557	-.2149862
_cons	3.018068	.0138501	217.91	0.000	2.990922	3.045214
var(e.gpa)	.0475351	.0024437			.0429789	.0525742
var(e.hsgpa)	.0602102	.0019041			.0565915	.0640602
corr(e.gra~e, e.gpa)	.2754647	.1003886	2.74	0.006	.0697401	.4587145
corr(e.hsgpa, e.gpa)	.1905273	.081385	2.34	0.019	.0273572	.3438079
corr(e.hsgpa, e.graduate)	.1534595	.1210009	1.27	0.205	-.0879677	.3778581
var(gpa[colle~d])	.0646465	.0097678			.0480764	.0869278
var(gra~e[col~d])	.9011305	.1745683			.6164413	1.317297
corr(gra~e[col~d], gpa[colle~d])	.2599483	.1069409	2.43	0.015	.0412395	.4548852

Now we see a random-effect variance parameter estimate for graduation and for college GPA and a correlation between these random effects. The student-level and college-level correlation parameters between the college GPA equation and graduation are significantly different from zero, so the researchers

conclude that there is endogenous sample selection. The student-level correlation between college GPA and high school GPA is also significantly different from zero, so they conclude that high school GPA is an endogenous covariate.

We can interpret the coefficients in the main equation as we did in [ERM] **Example 8a**, but now they are estimated for the population of admitted students, not the population of graduates. The estimated effect of high school GPA is slightly higher, 0.98 rather than 0.94.

Also see

[ERM] **eregress** — Extended linear regression

[ERM] **eregress postestimation** — Postestimation tools for eregress and xteregress

[ERM] **Intro 3** — Endogenous covariates features

[ERM] **Intro 4** — Endogenous sample-selection features

[ERM] **Intro 6** — Panel data and grouped data model features

[ERM] **Intro 9** — Conceptual introduction via worked example